Large Language Models Are Human-Level Prompt Engineers

Abstract

By conditioning on natural language instructions, large language models (LLMs) have displayed impressive capabilities as general-purpose computers. However, task performance depends significantly on the quality of the prompt used to steer the model, and most effective prompts have been handcrafted by humans. Inspired by classical program synthesis and the human approach to prompt engineering, we propose Automatic Prompt Engineer (APE) for automatic instruction generation and selection. In our method, we treat the instruction as the"program,"optimized by searching over a pool of instruction candidates proposed by an LLM in order to maximize a chosen score function. To evaluate the quality of the selected instruction, we evaluate the zero-shot performance of another LLM following the selected instruction. Experiments on 24 NLP tasks show that our automatically generated instructions outperform the prior LLM baseline by a large margin and achieve better or comparable performance to the instructions generated by human annotators on 19/24 tasks. We conduct extensive qualitative and quantitative analyses to explore the performance of APE. We show that APE-engineered prompts can be applied to steer models toward truthfulness and/or informativeness, as well as to improve few-shot learning performance by simply prepending them to standard in-context learning prompts. Please check out our webpage at https://sites.google.com/view/automatic-prompt-engineer.